Offline Handwritten Gujarati Word Recognition

Parita R. Paneri
Dharmsinh Desai University
Gujarat, India
Email: paritapaneri95@gmail.com

Ronit Narang
Dharmsinh Desai University
Gujarat, India
Email: ronitnarang11@gmail.com

Mukesh M.Goswami
Dharmsinh Desai University
Gujarat, India
Email: mukesh.goswami@gmail.com

***Abstract*—Gujarati language is an Indo-Aryan language that has a complex structure wherein extracting each character becomes hectic because of the presence of diacritics. Implementation of word recognition technique on the Gujarati database makes the work easy as it does not require extraction of symbols and glyphs from the word image. In this paper, we describe a handwritten Gujarati word recognition technique using Histogram of Oriented Gradients (HoG) features and state of the art classifier like Support Vector Machine (SVM) and k-Nearest Neighbor (kNN). The experiments were performed on a moderate sized database of handwritten Gujarati city names. The work produced has direct application in handwritten postal address processing.**

I. Introduction

Offline handwritten word recognition is a widely studied pattern recognition problem and has direct applications in automated cheque processing, handwritten postal mail sorting, automatic processing of handwritten forms etc. The problem can be solved using three possible approaches, namely incremental, holistic and hybrid[1]. In the incremental approach the word image is further divided into segments and uses incremental model for recognition whereas in the holistic approach the entire word is considered as a single unit of recognition. The hybrid approach is a mix of two. In this paper, holistic approach for word recognition has been proposed to identify handwritten city names in Gujarati script. Gujarati is 26th most-spoken native language in world with 65.5 million Gujarati speakers all over the world. The language also has a rich collection of literary work including the handwritten notes by M. K. Gandhi [2]. There are 13 vowels and 34 consonants (shown in Fig. 1). Apart from vowels and consonants, Gujarati word construction also contains diacritics (Matras) (shown in Fig. 2) because of which, character recognition in Gujarati language becomes difficult. This challenge can be addressed by recognizing the whole word. However, such system can only be used in the domain specific problems where the size of the vocabulary is limited. For example, automated postal address sorting using city names or automated processing of handwritten forms etc. The implementation of word recognition system includes pre- processing of word images, segmentation of words, feature extraction, and classification of word images.

The rest of the paper is organized as follows: Section II includes survey; Section III describes the proposed work. Section IV includes experimental results followed by conclusion in Section V.

Fig. 1. Gujarati Character Set

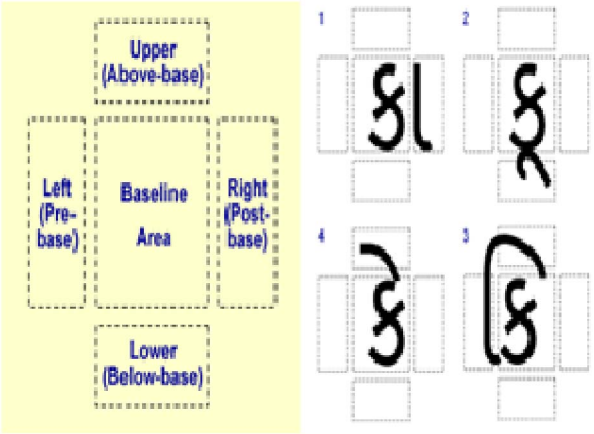


Fig. 2. Gujarati character with diacritics

II. Literature Review

Some work has been reported for handwritten word recognition methods for the dominating Indian scripts like Devanagari[1][3-4], Bengali[5], and Tamil[6].

In 2008, Bikas et al. have reported their work on handwritten word recognition for Devanagari script using both holistic as well as segmentation based approach. The holistic approach uses directional chain code and hidden markov model(HMM) classifier [3] whereas the segmentation based method divides word image into pseudo characters and extract stroke based features from each character. The classification algorithm used

978-1-5090-6734-3/17/$31.00 ©2017 IEEE

was HMM[4]. A huge database of size 22500 words was used in both the experiments. The accuracy reported in former case was 80.02% while in latter case it was 84.31%. Recently in 2016, Satish Kumar [1] has proposed a segmentation based technique for handwritten Devanagri word recognition. The experiments were performed on a database of size 3500 words and claimed 80.8% and 72.0% accuracy for two and six character words, respectively.

Bhowmik et al. [5] have put some efforts in recognition of handwritten Bangla word. They used holistic approach and recognized words from their overall shape described using Histogram of Oriented Gradients (HOG) features. The accuracy claimed was 87.35% using a database of 1020 samples using MLP classifier.

Subramaniam. T et al. [6] have mentioned a novel approach for Tamil word recognition. Most of the Tamil words are touching and overlapping. To overcome the hurdle they employed Gabor based features to process them followed by SVM classifier for recognition. They reported 86.36% recognition rate on a database of 4270 samples of 217 country names.

Some work was reported in the literature for offline Gujarati handwritten numeral [7] and symbols recognition [8]. However, to the best of our knowledge, there is no work reported for handwritten Gujarati word recognition so far. Therefore, the efforts in direction are justified.

III. Proposed Approach

The proposed method consists of three steps, namely database collection, feature extraction, and classification as shown in Fig.3 In the first step, a database of 2700 handwritten Gujarati city name samples was created since there is no such database exist in the literature. The HoG features were extracted from each word image and finally, the samples were classified using kNN and SVM classifier.

A. Database Collection

Since there does not exist any Gujarati handwritten word image database in public domain, a small database was generated in house for experimental purpose. The database consists of 2700 samples of handwritten Gujarati city name of 10 well-known cities collected from 65 different subjects. The subjects were selected from different age, gender, and professional background like teachers, professors, students, admin staff, peons as well as random subjects. Each subject was asked to write on the data-sheet (shown in Fig. 4) using two different pens. This ensures the verities of writing style and ink thickness in samples.

The data-sheets were scanned using 300 dpi flat-bed scanner and converted into digital images. The digitized images are binarized using Otsu's method [9] which uses a global threshold determined dynamically from the image. The choice seems reasonable as the flat-bed scanner has a uniform lighting condition. The binarized document images are prone to the salt and pepper noise during acquisition or transformation. Therefore, median filter method is employed next to denoise the images before segmentation. Image segmentation is the

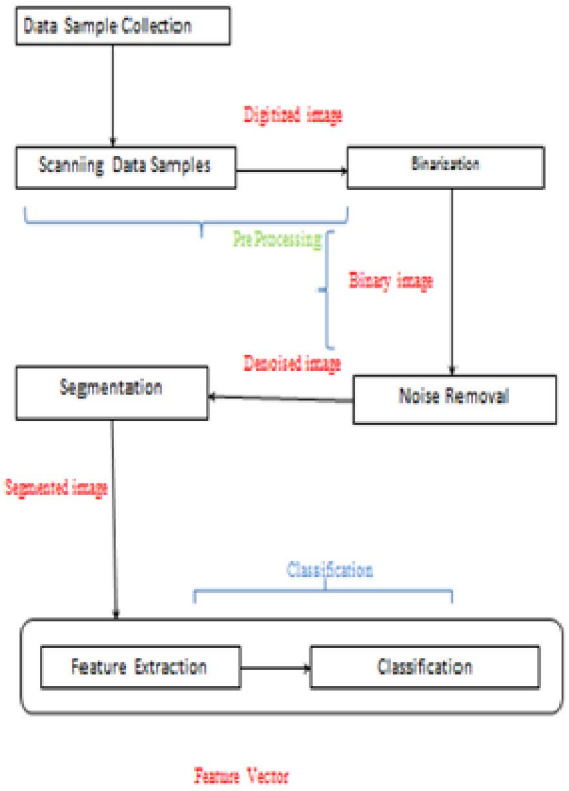


Fig. 3. Proposed Approach

process of partitioning a digital image into multiple segments (sets of pixels, also known as super-pixels). Each segment represents a region of interest to be analyzed further. For word recognition task, the document images are segmented in lines and words. Horizontal and vertical projection profiles [10] are used for line and word segmentation, respectively(as shown in Fig. 5 and 6).

B. Feature Extraction

The feature extraction starts from an initial set of measured values and builds derived values intended to be informative, non-redundant, and should facilitating the subsequent learning.

The histogram of oriented gradients (HOG) [11] is a feature descriptor used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image and hence describes local shape of an object. The histogram is computed for each of the dense grid of uniformly spaced and non-overlapping cells.

C. Classification

In the proposed work, the base line experiment was performed using k-NN classifier and the results were further improved using more advance SVM classifier

The k-Nearest-Neighbor (k-NN) is a popular non-parametric recognition method, where a posteriori probability is estimated from the frequency of nearest neighbors of the unknown

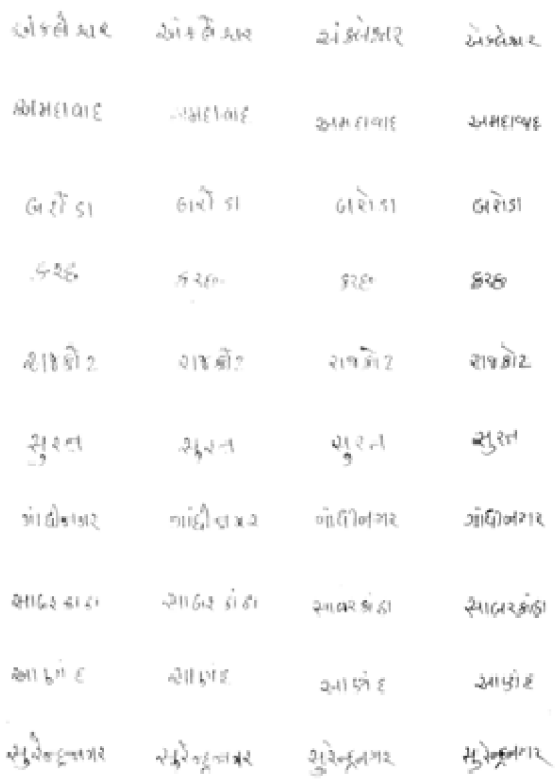


Fig. 4. Sample data-sheet for handwritten Gujarati word database

Fig. 5. Segmented Line Image

pattern. Advantages of this method are: it can handle large number of classes; it can avoid over fitting of parameters and requires no learning or training phase. Compelling recognition results were reported using this approach particularly for handwritten recognition. The drawback of this method is the high computational cost when the classification is conducted. After computing HoG feature vector we fed them to K-nearest neighbor classifier for recognition. Where K is number of maximum neighbors and that will used to find nearest neighbor. We choose k=3 as we are having 10 different city names i.e. class-labels (k = Sqrt(10)) [12].

Support Vector Machine (SVM) is based on the statistical learning theory and quadratic programming optimization. The

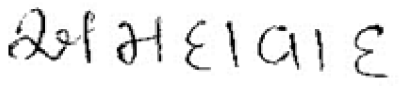


Fig. 6. Segmented Word Image

learning algorithm in SVM finds the hyper plane which is at a maximum distance form support vectors [13]. It is a binary classifier but multiple SVMs can be combined to form a system for multi-class classification. The performance of SVM depends on the type of kernel function used. The kernel functions are employed to project the low-dimensional data into high-dimensional projection space. In our experiment linear kernel function is used since the size of the feature vector is already large as compare to the number of classes.

IV. Experiments and Results

TABLE I
CLASS LABEL. CITY NAME. AND WORD IMAGE

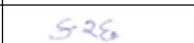
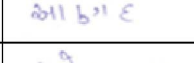
Sr. No	Class Label	City Name	Gujarati Word Image
1	ANK	Ankleshwar	
2	AHM	Ahmedabad	
3	BRD	Vadodara	
4	KCH	Kutch	
5	RAJ	Rajkot	
6	SRT	Surat	
7	GDN	Gandhinagar	
8	SAB	Sabarkantha	
9	AND	Anand	
10	SRN	Surendranagar	

TABLE II
ACCURACY OBTAINED ON HANDWRITTEN GUJARATI WORD DATABASE USING K-NN AND SVM

	k-NN	SVM
Gujarati Word Image Database	76.87%	85.87%

The proposed method was implemented in Python using OpenCV library [14]. Segmented word-images are resized to 200x100 sizes before compute the feature vector. The image is further divided into 32 non-overlapping blocks (cells) each of size 25x25 and the histogram of 9 different gradients were computed in each block. The computed histogram is further normalized using block wise normalization method and concatenated together to form the feature vector. Thus, the size of the feature vector is 289 where 288 is the HoG feature

TABLE III
COMPARATIVE ANALYSIS OF PROPOSED WORK WITH EXISTING WORK

Script	Reference	Feature Used	Classifier	Database	Accuracy
Devanagari	Bikas et al[3]	Directional Chain Code	HMM	22500	80.02%
	Bikas et al[3]	Stroke Feature	HMM	22500	84.31%
	Kumar S.[1]	Multiple features like	MLP	3500	72.0-80.8%
		Neighbor Pixel Weight, Vertical and Horizontal histogram, Crossover points, Gradients			
Bengali	Bhowmik et al.[5]	HoG	MLP	1020	87.35%
Tamil	Subramaniam. T et al.[6]	Gabor based Features	SVM	4270	86.36%
Gujarati	Proposed Work	HoG	SVM	2700	85.87%

TABLE IV
CONFUSION MATRIX

	ANK	AHM	BRD	KCH	RAJ	SRT	GDN	SAB	AND	SRN	
ANK	185	5	0	2	4	1	5	2	1	9	86.4
AHM	7	187	0	0	0	1	7	3	10	1	86.6
BRD	4	0	175	1	1	7	2	0	1	2	86.1
KCH	8	0	2	190	0	9	0	4	1	1	88.4
RAJ	5	6	5	0	183	0	2	1	8	3	85.9
SRT	5	1	6	6	1	193	0	1	1	4	89.3
GDN	11	10	0	0	1	2	181	3	1	6	84.2
SAB	2	6	0	5	4	0	3	179	4	3	86.9
AND	3	17	4	1	10	0	1	3	170	2	84.6
SRN	15	8	0	0	2	3	7	5	2	171	80.3

Average Accuracy is **85.8**

vector (i.e. 9x32 = 288) and the last column indicates class label. In our case, there are 10 different class labels shown in Table 1. The average test accuracy obtained using 10-fold cross validation is used as a primary performance measure. (i.e. the database is divided into 10 folds and the experiment is repeated 10 times with one fold used for testing and remaining 9 fold for training in round robin order to estimate the average test accuracy)

The average test accuracy obtained on word-image database using kNN and SVM is 76.78% and 85.87%, respectively(as shown in Table 2). Thus, it is evident that SVM gives significant gain in accuracy compare to kNN. The detailed confusion matrix in Table 4 shows that SRN being the most frequently misclassified class label followed by GDN and AND, respectively. It is because of unconstrained writing style. We can resolve these confusions by applying some additional heuristics. The comparative summary of the proposed work with other existing similar approaches (shown in Table 3) indicates that the result obtained are comparable with other existing work.

V. CONCLUSION

The paper presents the first ever attempt for handwritten Gujarati word recognition. A small handwritten Gujarati city name database was created for experimental purpose and could be useful to other researchers in the field. The experiment reported a highest test accuracy of 85.87% using HoG feature and SVM classifier which is comparable with other existing work. The results could be improved further by using multiple features or additional heuristics to handle misclassified samples.

REFERENCES

[1] S. Kumar, "A study for handwritten Devanagari word recognition" *In Communication and Signal Processing (ICCSP), 2016 International Conference on*, pp. 1009-1014. IEEE, 2016.

[2] M. K. Gandhi, The Collected Works of Mahatma Gandhi, *1st ed. Min of Information and Broadcasting*, GoI, Vol. 1,1958

[3] S. Bikash, S. K. Parui and M. Shridhar. "Offline Handwritten Devanagari Word Recognition: A holistic approach based on directional chain code feature and HMM." *In Information Technology, 2008. ICIT'08. International Conference on*, pp. 203-208. IEEE, 2008

[4] S. Bikash, S. K. Parui and M. Shridhar, "Offline handwritten Devanagari word recognition: A segmentation based approach." *In Pattern Recognition, 2008.ICPR 2008. 19th International Conference on*, pp. 1-4. IEEE, 2008.

[5] S. Bhowmik, M. G. Roushan, R. Sarkar, M. Nasipuri, S. Polley and S. Malakar, Handwritten bangla word recognition using hog descriptor, *In Emerging Applications of Information Technology (EAIT) 2014 Fourth International Conference on*, pp. 193-197, IEEE, 2014.

[6] T. Subramaniam, U. Pal, H. Premaretne and N. Kodikara, Holistic recognition of handwritten Tamil words, *In Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on*, pp. 165-169, IEEE, 2012.

[7] M. M. Goswami and S. K. Mitra, Offline handwritten Gujarati numeral recognition using Low-Level Stroke, *Int. J. of Applied Pattern Recognition (IJAPR)*, Vol, 2(4), pp-353-379, 2015.

[8] S. J. Macwan and A. N. Vyas, Classification of offline Gujarati handwritten characters, *Advance in Computing, Comm. and Informatics (ICACCI15), International Conference on*, pp-1535-1541, IEEE, 2015.

[9] N. Otsu, A thresholding selection method for grey-level histogram *IEEE Trans. System, Man and Cybernetics*, Vol. 9, pp-62-66, IEEE, 1979.

[10] L. O'Gorman and K. Rangachar, *Document image analysis*, Vol. 39. Los Alamitos: IEEE Computer Society Press, 1995.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection" *In Computer Vision and Pattern Recognition, CVPR 2005, IEEE Computer Society Conference on*, vol. 1, pp. 886-893. IEEE, 2005.

[12] O. Boiman, E. Shechtman and M. Irani, "In defense of nearest-neighbor based image classification" *In Computer Vision and Pattern Recognition, 2008. CVPR 2008 IEEE Conference on*, pp. 1-8. IEEE, 2008.

[13] C. Cortes and V. Vapnik, "Support vector machine" *Machine learning*, Vol. 20, no. 3, pp.273-297, 1995.

[14] I. Culjak, D. Abram, T. Pribanic, H. Dzapo and M. Cifrek, "A brief introduction to OpenCV", *In MIPRO, 2012 proceedings of the 35th International Convention*, pp. 1725-1730. IEEE, 2012.